



BITS Pilani
DIGITAL
Excellence made yours

BILL SCANNING BASED REWARD & INTELLIGENCE SYSTEM

Proposal Report

The Driving Force Behind

TEAM ARAJ

Group 120

Ashfaaq

Ranjeet

Arpan

Jyoti

02.FEB.2026

Arp

PROJECT DETAILS



Course Title **B.Sc. Computer Science**
BITS Pilani Digital

Project Title **Bill Scanning Based Reward & Intelligence System**
AI/ML

Group No. **120**

Group Name **TEAM ARAJ**

Number of
members **4**

Members	A shfaaq Feroz Muhammad	2023EBCS005
	R anjeet Singh	2023EBCS457
	A rpan Chatterjee	2023EBCS399
	J yoti Kataria	2023EBCS435

Group
Advisor **Prof. Uma Sankara Rao**

Submission
Date **02 | February | 2026**

INDEX



No	Title		
1	Cover Page - Project Title	14	System Architecture
2	Project Details	15	ML Pipeline
3	Index page	16	Proposed Model Types
4	Team Details	17	Deployment Model
5	Advisor Details	18	Dataset
6	Project Overview	19	Evaluation Metrics
7	Problem statment	20	Flow Diagram
8	Motivation	21	Objectives
9	Background Study	22	Scope & Constraints
10	Identified Gap	23	Conclusion - End
11	Industry Systems		
12	Existing System Analysis		
12	Existing System Analysis		
13	Solution Statement		

TEAM ARAJ



Who We Are..!

Team **ARAJ**, **Group 120**, is a four-member student team from the **B.Sc. Computer Science program (BITS Pilani Digital)**. This study project is completed as part of our final-year academic requirements, emphasizing real-world problem identification and the creation of feasible, technology-driven solutions.

Our chosen project area is **Artificial Intelligence and Machine Learning (AI/ML)**, focusing on data extraction, pattern recognition, and converting unstructured data into meaningful, privacy-compliant intelligence.

Team Members



Ashfaaq Feroz Muhammad
2023EBCS005

B.Sc. Computer Science student interested in blending creativity with technology, focusing on AI/ML-driven system design, problem identification, and innovative solution development.



Ranjeet Singh
2023EBCS457

B.Sc. Computer Science student focused on analytical thinking, feasibility analysis, and understanding real-world system challenges for practical AI/ML-based solutions.



Arpan Chatterjee
2023EBCS399

B.Sc. Computer Science student with a strong hold in marketing. Scaling User Acquisition for apps at MillionTap. Problem solving attitude with literacy in AI-ML based solutions and builder mentality.



Jyoti Kataria
2023EBCS435

B.Sc. Computer Science student focused on user-centric analysis, clear communication, documentation, and presenting AI/ML solutions in an understandable manner.

OUR ADVISOR



Prof. Uma Sankara Rao

umasankara.rao@online.bits-pilani.ac.in

A PhD-qualified researcher specializing in cloud security, blockchain, AI, ML, and emerging technologies, committed to advancing innovation through impactful, industry-focused research and development.

OVERVIEW

BILL SCANNING BASED REWARD & INTELLIGENCE SYSTEM



The proposed solution is a **bill scanning-based reward and intelligence system** that converts everyday purchase receipts into mutual value for users and brands. Users simply scan bills from any category—such as groceries, travel, utilities, or electronics—and receive instant rewards in the form of cashback or coupons. Beyond rewards, the system extracts structured data from bills and transforms it into privacy-compliant, anonymized purchase-level intelligence.

This intelligence layer enables analysis of spending patterns, brand preferences, category-wise trends, and value segmentation without exposing individual user identities. Unlike traditional loyalty or cashback platforms that operate in silos, the solution works across merchants and categories. It is designed as an SDK-based technology, allowing seamless integration into existing or new applications, ensuring scalability and wide distribution.

Overall, the system unlocks hidden value from unutilized bill data, incentivizes user participation through rewards, and provides brands with actionable insights to deliver more relevant offers and partnerships.

PROBLEM STATEMENT

Consumers produce vast amounts of transactional data through their everyday bills and receipts, spanning categories like groceries, utilities, travel, electronics, and services. Unfortunately, this data is often fragmented, unstructured, and left unused after the transaction concludes. Meanwhile, brands do not have access to valuable cross-platform, purchase-level intelligence that reflects actual consumer behavior outside their own ecosystems.

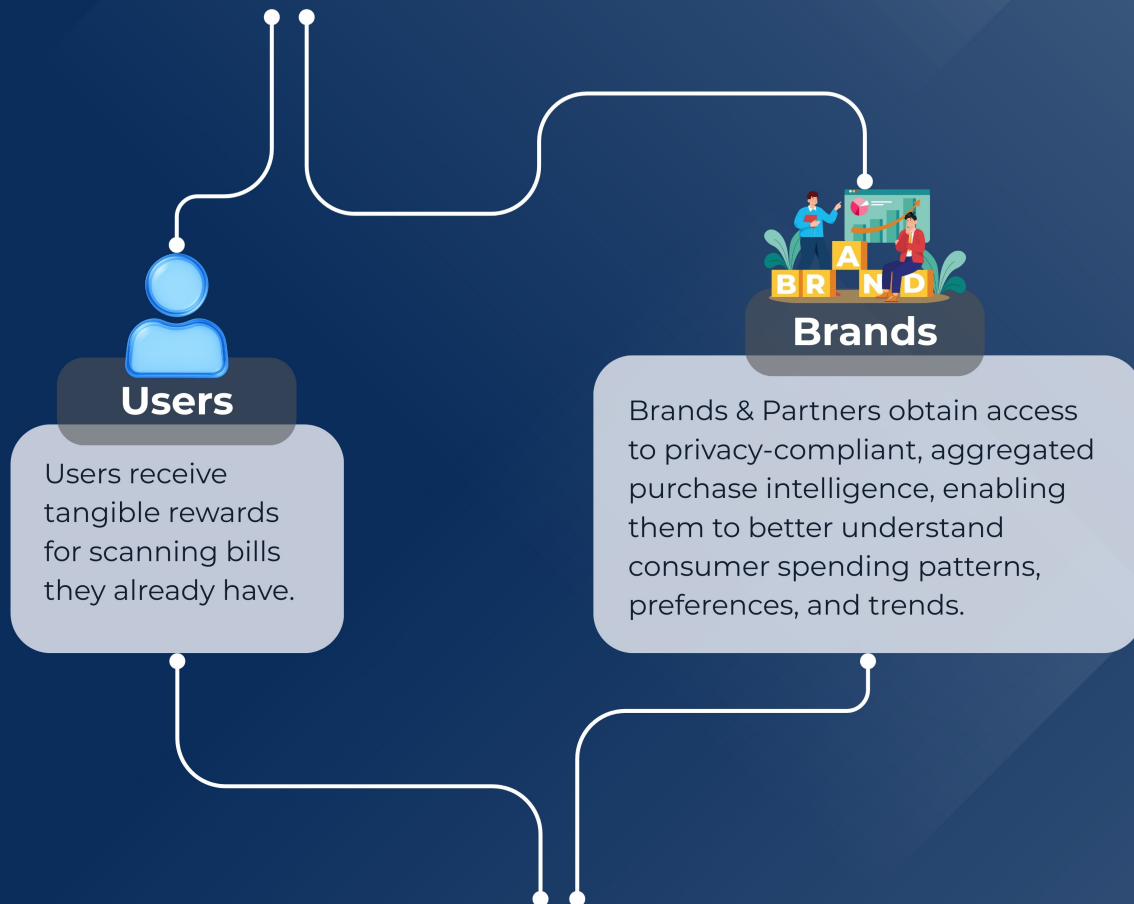
Existing reward and cashback systems are typically brand-specific, siloed, and limited in the insights they provide. These systems fail to capture a comprehensive view of consumer spending across different merchants, categories, and time periods.

Summary

Consumers generate valuable bill data that remains fragmented and unused, while brands lack cross-platform purchase insights. Existing reward systems are siloed, brand-specific, and fail to capture holistic consumer spending behavior.

MOTIVATION

The primary motivation for this project is to design a mutually beneficial system where:



By transforming raw bill images into anonymized behavioral insights, the system unlocks value from data that would otherwise be wasted.

LITERATURE REVIEW & BACKGROUND STUDY

Receipt Scanning & OCR

Advances in **Optical Character Recognition (OCR)** and computer vision have made it possible to automatically extract structured data from invoices and receipts. Tools such as expense trackers and accounting applications utilize OCR for digitizing bills, but their focus remains on individual record keeping rather than producing large scale intelligence.

Consumer Analytics & Behavioral Data

Modern consumer analytics primarily relies on first-party and third-party data, which is often restricted to behavior observed within specific platforms, such as E-commerce browsing. Access to cross-merchant, purchase level data is rare, expensive, and typically gathered through surveys or panel data, limiting both accuracy and scale.

Reward-Driven Data Sharing

Research indicates that users are more willing to share personal data when they perceive a clear value exchange, such as cashback, discounts, or loyalty rewards. Incentivized data contribution models have demonstrated success in areas like fintech and survey platforms.

IDENTIFIED GAP

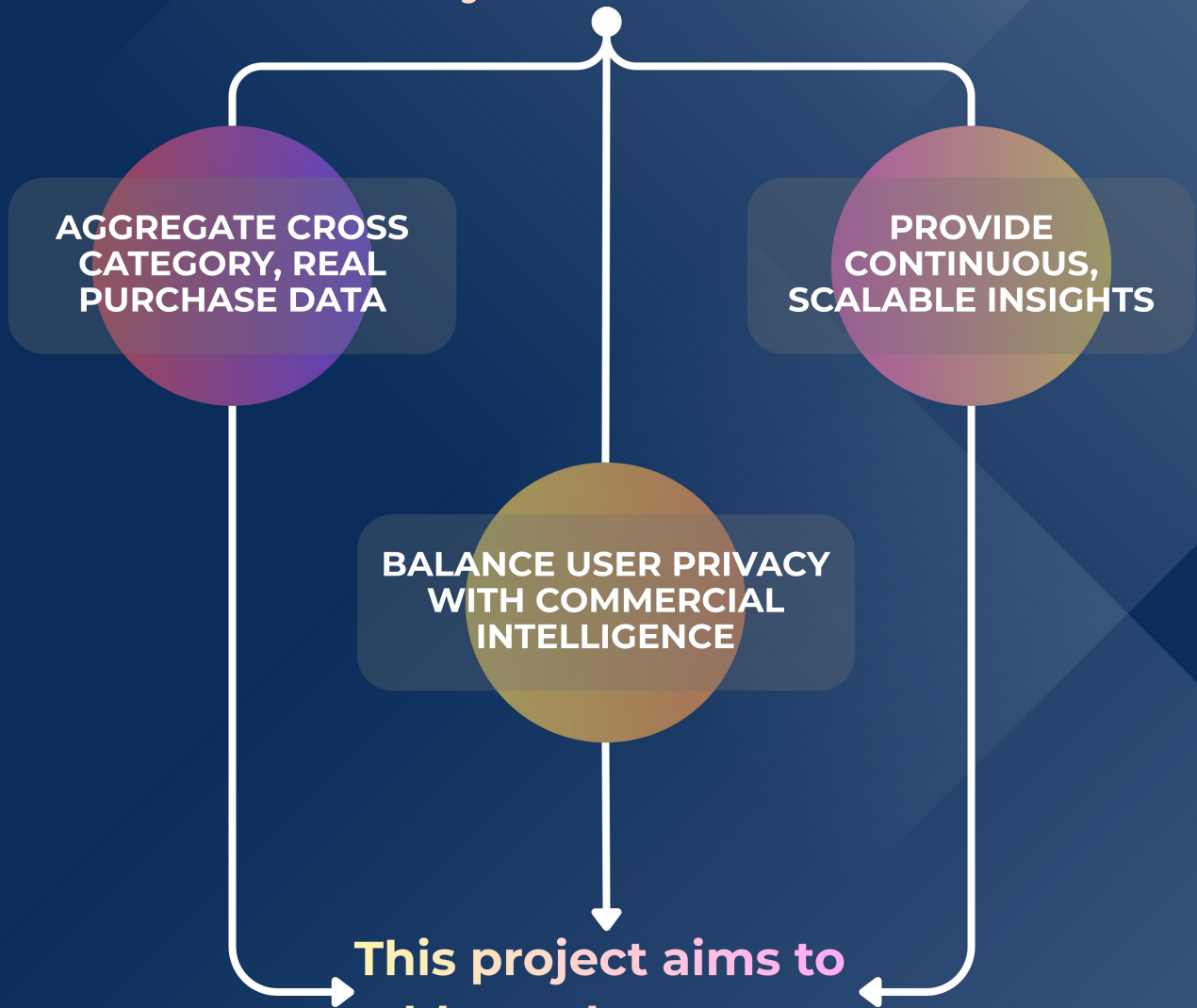
Current systems are unable to:

AGGREGATE CROSS
CATEGORY, REAL
PURCHASE DATA

PROVIDE
CONTINUOUS,
SCALABLE INSIGHTS

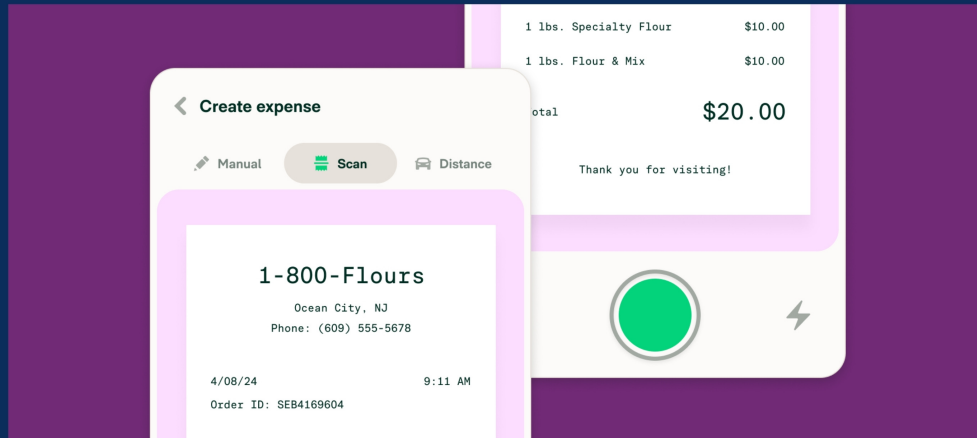
BALANCE USER PRIVACY
WITH COMMERCIAL
INTELLIGENCE

This project aims to
address these gaps.



COMMERCIAL & INDUSTRY SYSTEMS

Expensify SmartScan



A popular receipt scanning workflow where users capture receipts and the system extracts expense details automatically. This demonstrates real-world viability of OCR-based expense extraction pipelines.

Amazon Textract – Analyze Expense



A dedicated API for invoices/receipts that returns structured JSON including summary fields and line item groups. Shows a production-grade approach to receipt understanding beyond plain OCR.

Microsoft Azure Document Intelligence



A prebuilt receipt model combining OCR and deep learning to extract key receipt fields. Demonstrates strong baseline capability for field extraction in varied receipt formats.

RESEARCH SYSTEMS & ACADEMIC FOUNDATIONS

.....

SROIE (ICDAR 2019 Competition)

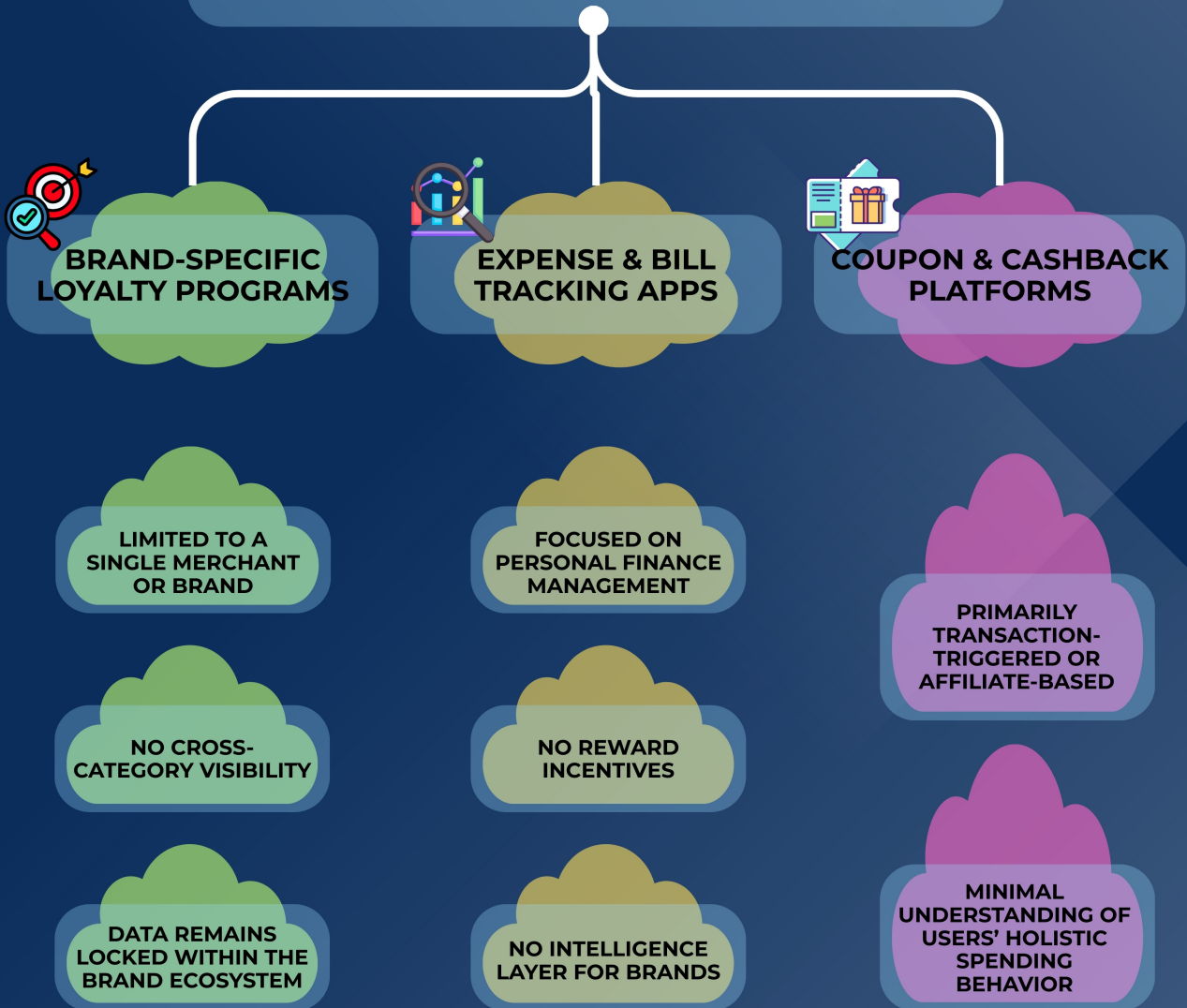
Establishes standard tasks for receipt OCR and key information extraction and provides evaluation protocols, making it a suitable benchmark for academic validation.

Donut (OCR-free Document Understanding Transformer)

Proposes an OCR-free end-to-end architecture that avoids OCR error propagation and can improve robustness in structured extraction tasks; suitable as a future extension beyond Phase-1.

EXISTING SYSTEM ANALYSIS

Existing Approaches



Limitations of Existing Systems



SOLUTION STATEMENT

The proposed system allows users to **scan bills from any merchant and receive instant rewards** such as cashback, coupons, or offers.

Simultaneously, the platform converts the extracted data into privacy-compliant, anonymized intelligence.

In addition to physical bills, users also generate significant amounts of transactional data through SMS notifications and OTP messages.

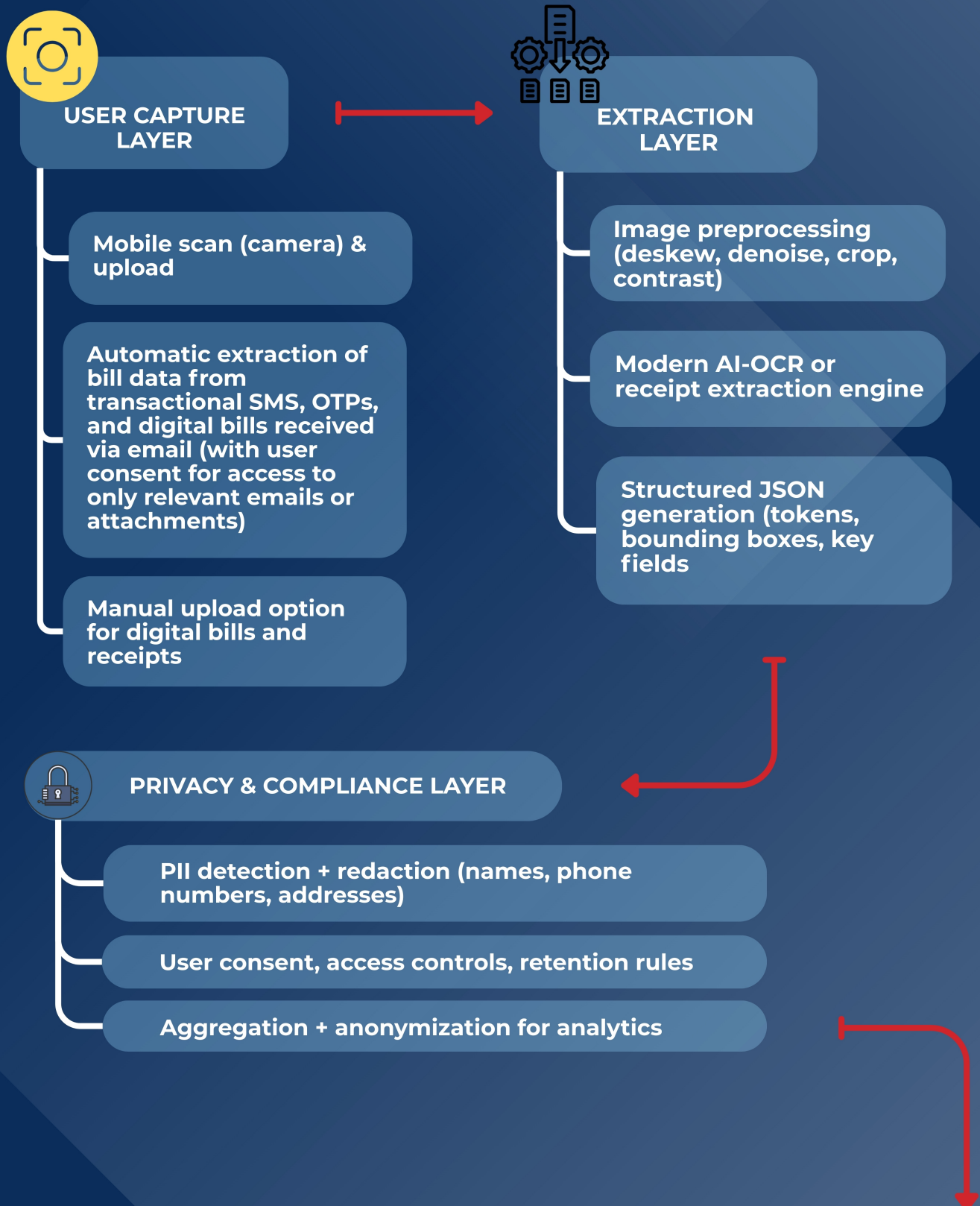
Furthermore, **users frequently receive digital bills and receipts via email.**

With user consent, the system can be granted automated access—restricted solely to emails with bill-related subjects or PDF receipts—for seamless extraction of digital bill data. Alternatively, users may choose to manually upload digital bills, ensuring flexibility and control over their data sharing preferences.

Summary

The system enables users to scan physical or digital bills to earn instant rewards, while converting data into privacy-compliant intelligence. With user consent, bill data can be extracted from SMS, emails, or manual uploads.

SYSTEM ARCHITECTURE





ML INTELLIGENCE LAYER

Category classification

Brand/Merchant affinity and category preferences

Insight computation (spend trends, affinity scores)

Basket/line-item parsing



REWARD ENGINE

Reward rules (category-based / value-based)

Fraud checks (duplicate receipts, reuploads)

Wallet / coupon issuance



ANALYTICS & BRAND INSIGHTS PORTAL

Aggregated, privacy-compliant dashboards

Category trends, share-of-wallet proxies, cohort insights

ML PIPELINE

OCR → Cleaning → Categorization → Insights

.....

OCR / Receipt Extraction

Extract text blocks, words, bounding boxes, and confidence scores.

If using managed services:

- **AWS Textract AnalyzeExpense** is designed for invoices/receipts and returns structured expense objects.
- **Azure Document Intelligence** Receipt model extracts receipt fields (merchant name, date, total, tax) as structured JSON.
- **Google Vision OCR** provides text detection/extraction on images.

Cleaning & Normalization

Fix OCR artifacts:

currency symbols, decimal separators, merged tokens, date formats.

Normalize merchants using:

Regex cleanup + fuzzy match + embedding similarity

Validate totals:

reconcile subtotal + tax $\approx$ total where possible (tolerance threshold).

Categorization & Parsing

Extract features:

merchant string, item keywords, quantities/prices, inferred category terms

Model options:

- **Baseline:** rules + dictionaries (fast, interpretable)
- **ML:** supervised classifier for category prediction
- **KIE** for line-items and key-value fields (SROIE/CORD-aligned).

Insights Generation

Per-user analytics (private view):

monthly spend by category, merchant preference, top items

Aggregated intelligence (anonymized):

category-level trends, brand affinity, cohort segmentation

Predictive layer (Optional later):

churn/retention likelihood, offer response propensity.

PROPOSED MODEL TYPES

DOCUMENT TEXT EXTRACTION (OCR / TEXT DETECTION)

OPTION 1

Baseline, Fast to integrate

Use a mature OCR API (Google Vision / AWS Textract / Azure Document Intelligence). These services support robust OCR and, in some cases, receipt-specific field extraction.

OPTION 2

Research-grade, offline/controllable

Use an open-source OCR stack (e.g., text detection + recognition), or move toward OCR-free document understanding.

KEY INFORMATION EXTRACTION (KIE) / RECEIPT PARSING

PATH (1) POST-OCR KIE

Feasibility

Input: OCR tokens + bounding boxes

Model: Layout-aware Transformer for entity extraction (merchant, date, total, tax, line-items)

Dataset alignment: CORD (receipt parsing) and SROIE (receipt OCR + key field extraction).

PATH (2) OCR-FREE KIE

Advanced Extension

Model: Donut (OCR-free Document Understanding Transformer) for end-to-end structured extraction directly from the receipt image.

MERCHANT / CATEGORY / SPEND CLASSIFICATION

MERCHANT NORMALIZATION:

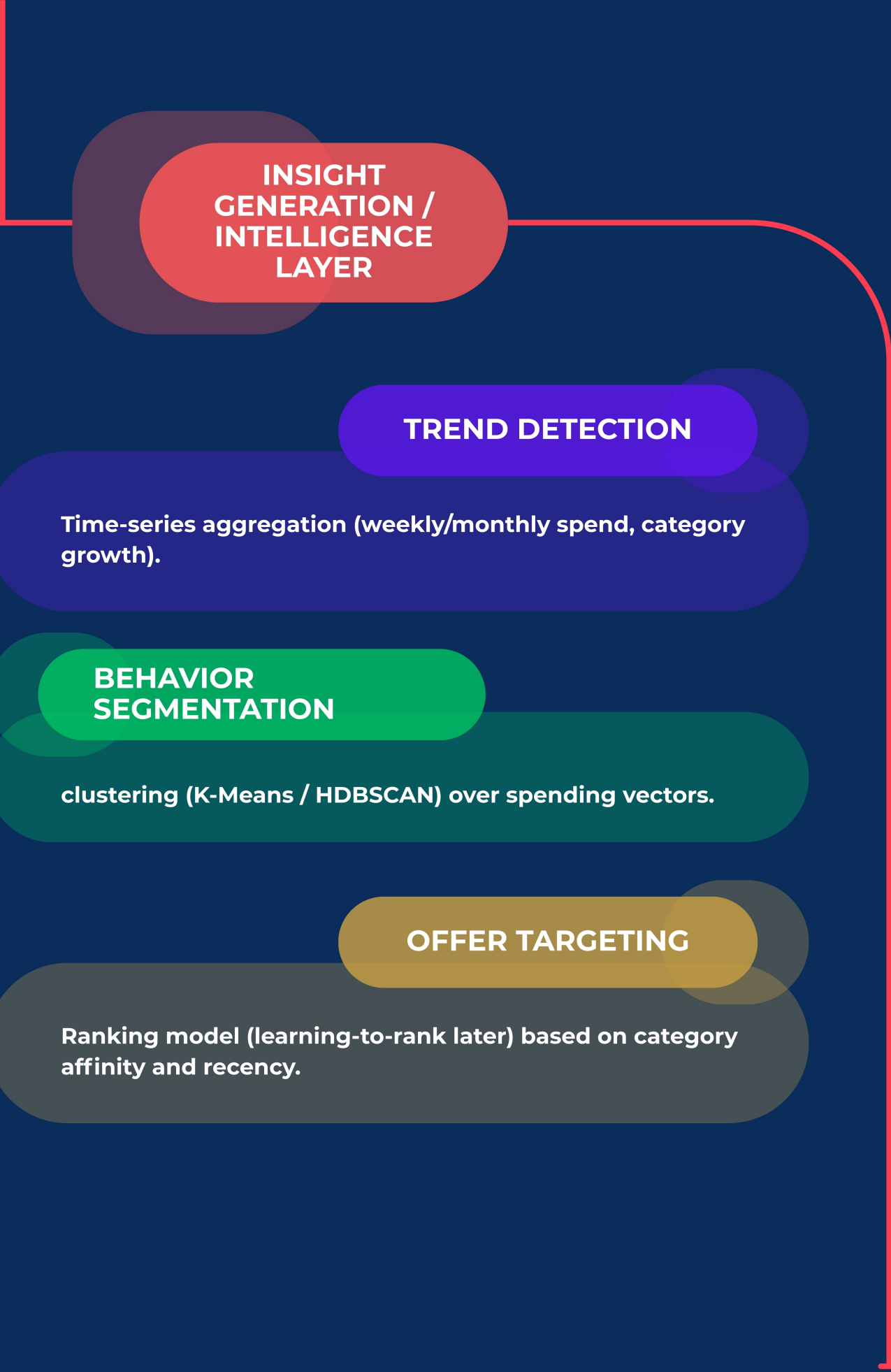
fuzzy matching + embedding similarity (e.g., sentence embeddings on merchant strings).

CATEGORY CLASSIFICATION

Groceries, Utilities, Electronics, etc.

Baseline: rules + keyword mapping

ML: supervised classifier (Logistic Regression / XGBoost) over engineered features (merchant, items, MCC-like patterns) and/or embeddings.



INSIGHT GENERATION / INTELLIGENCE LAYER

TREND DETECTION

Time-series aggregation (weekly/monthly spend, category growth).

BEHAVIOR SEGMENTATION

clustering (K-Means / HDBSCAN) over spending vectors.

OFFER TARGETING

Ranking model (learning-to-rank later) based on category affinity and recency.

DEPLOYMENT MODEL

OFFERED AS



Software
Development
Kit



Application
Programming
Interface

INTEGRATES WITH



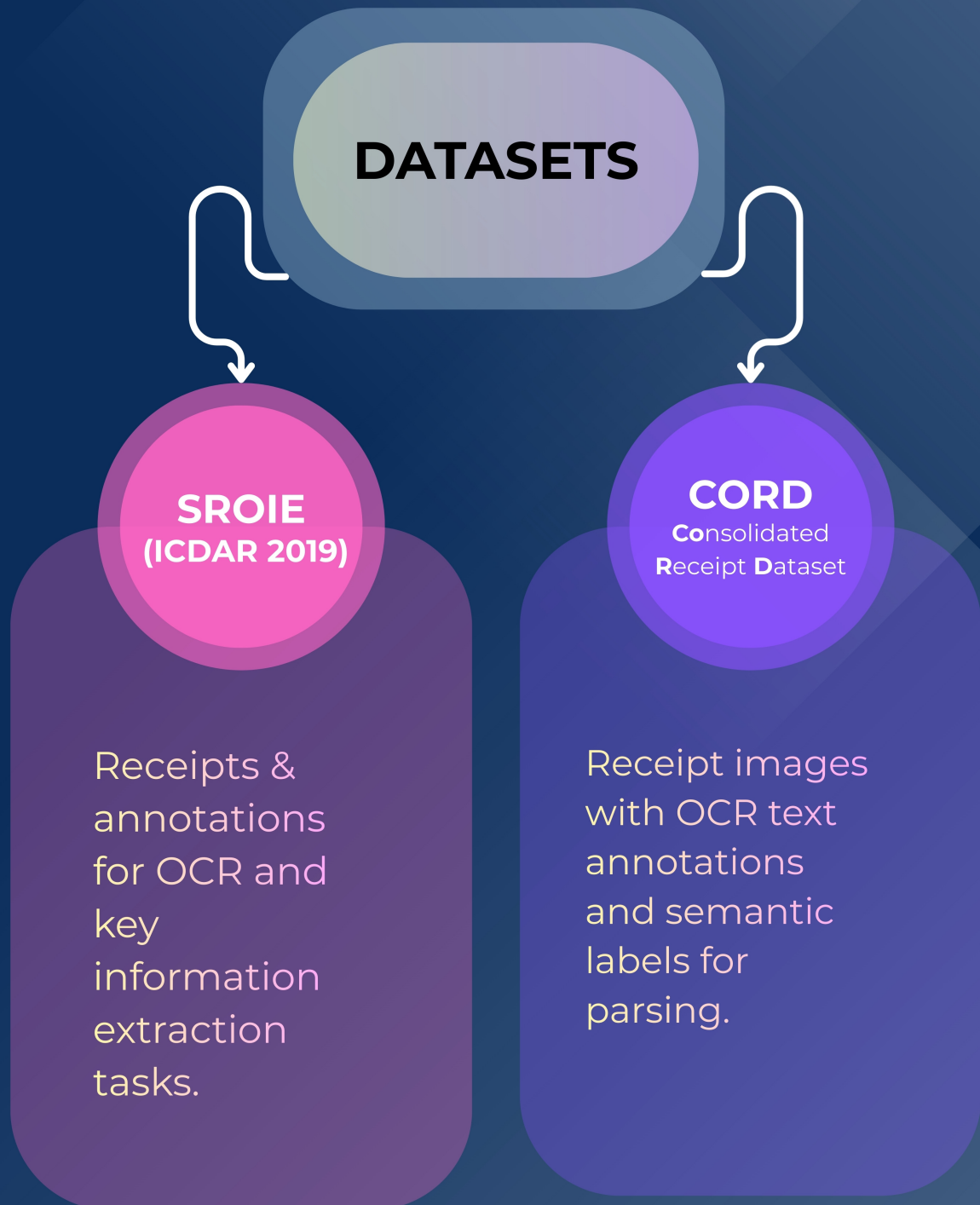
Existing
consumer apps

Fintech, wallet,
or loyalty
platforms

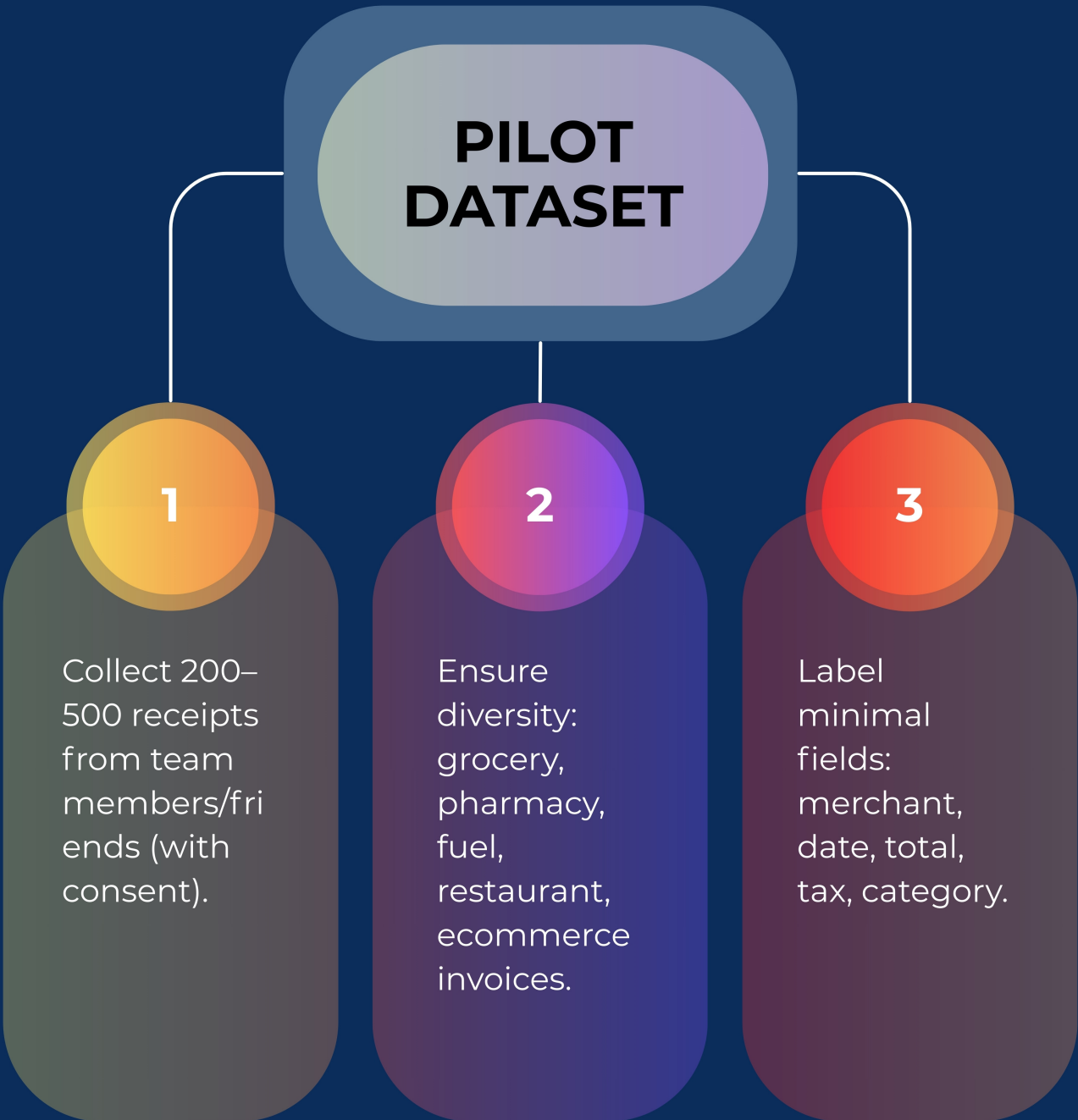


New standalone
implementations

DATASET



PILOT DATASET



```
graph TD; A[PILOT DATASET] --- B((1)); A --- C((2)); A --- D((3)); B --- E[Collect 200-500 receipts from team members/friends (with consent).]; C --- F[Ensure diversity: grocery, pharmacy, fuel, restaurant, ecommerce invoices.]; D --- G[Label minimal fields: merchant, date, total, tax, category.]
```

1

Collect 200–500 receipts from team members/friends (with consent).

2

Ensure diversity: grocery, pharmacy, fuel, restaurant, ecommerce invoices.

3

Label minimal fields: merchant, date, total, tax, category.

- Character Error Rate (CER), Word Error Rate (WER)
- Field-level extraction accuracy for key fields (merchant/date/total)

OCR Quality

- Entity-level Precision, Recall, F1 (for merchant/date/total/tax)
- Line item extraction: F1 on (item_name, price) pairs

KIE Key Information Extraction

EVALUATION METRICS

Category Classification

- Accuracy + Macro-F1 (to avoid bias toward frequent categories)
- Confusion matrix analysis (e.g., groceries vs household)

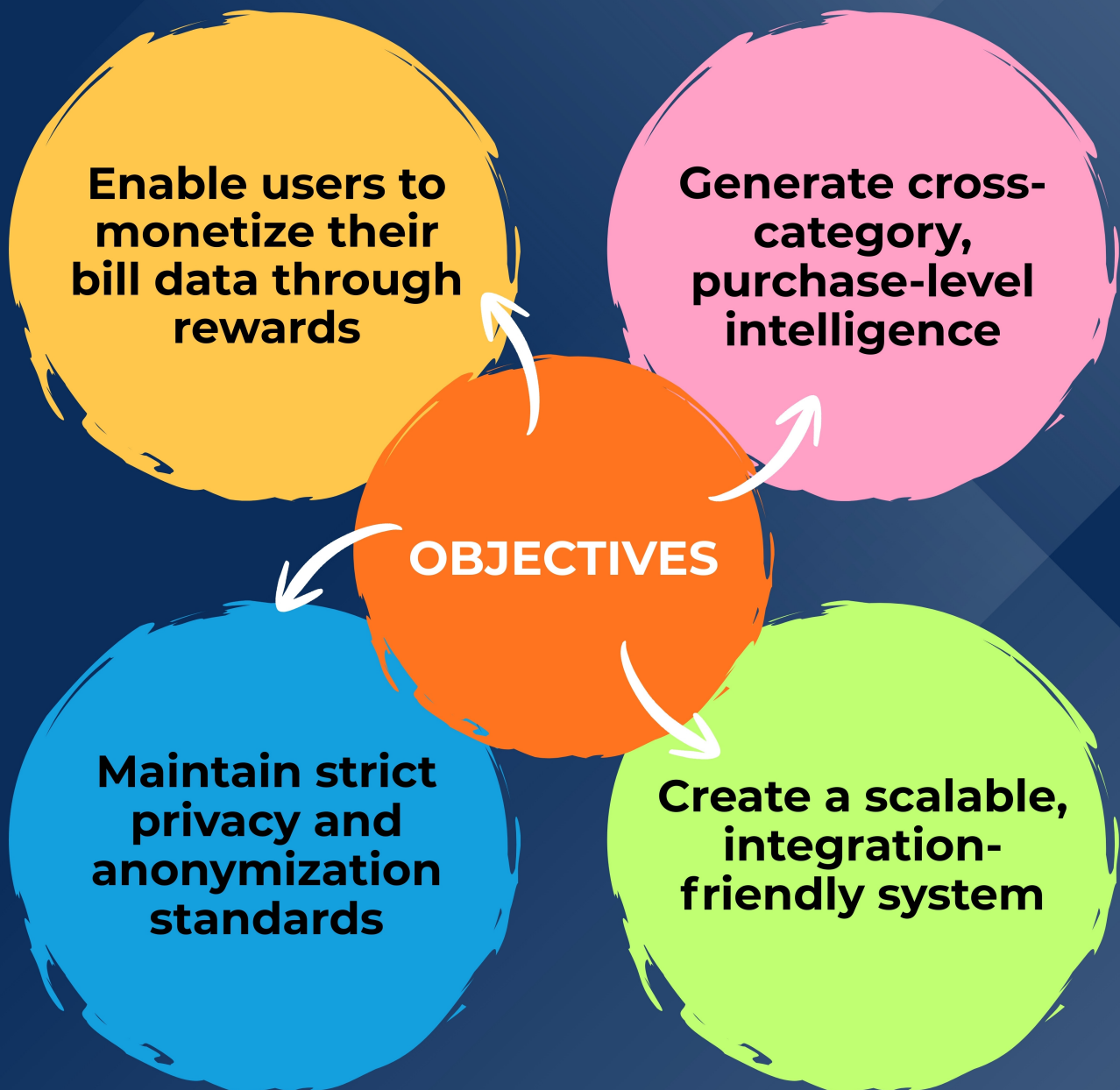
Insights Correctness

- **Aggregation stability:** compare computed totals vs extracted totals
- **Duplicate Detection Rate:** percent of repeated receipts correctly flagged

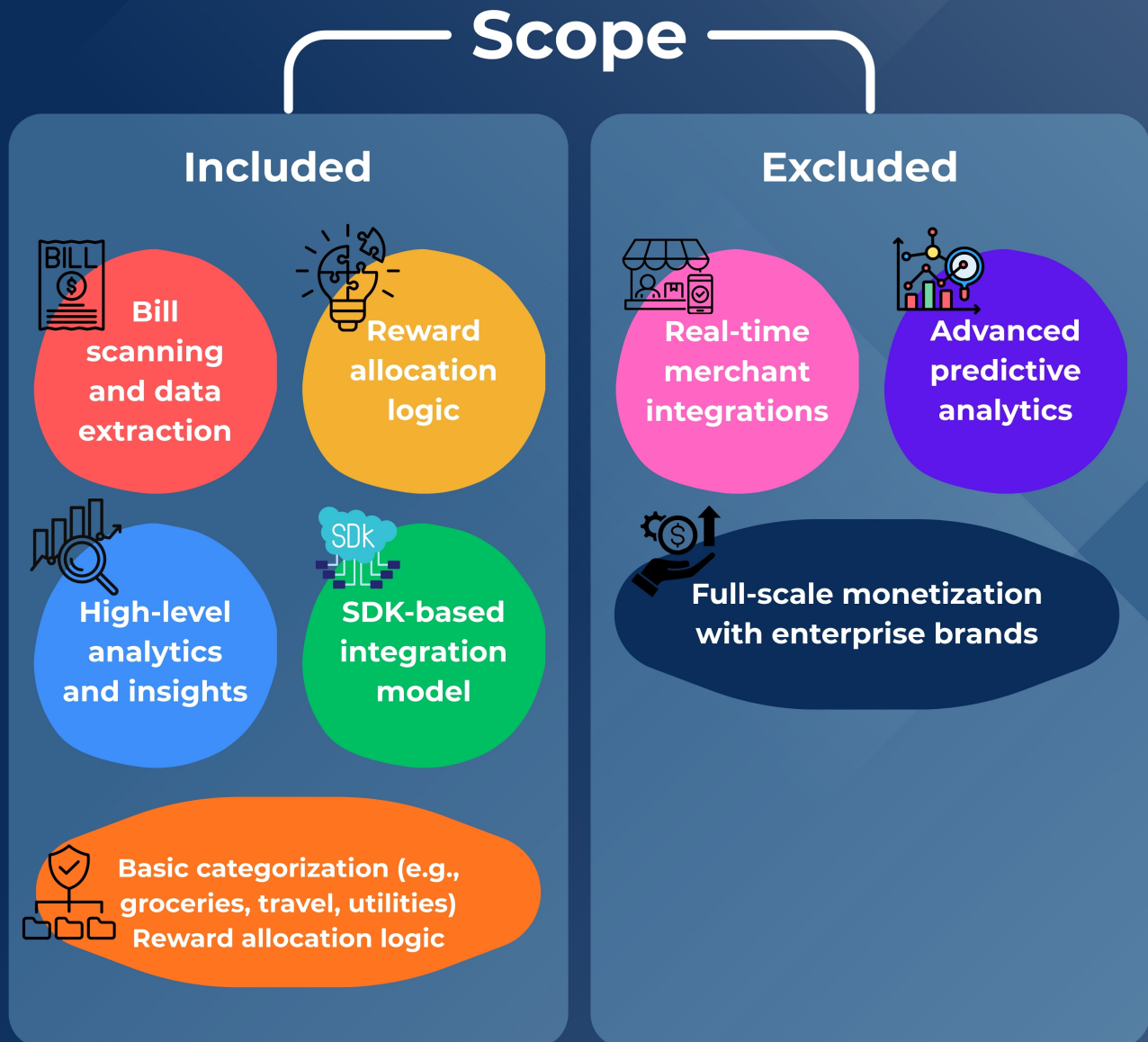
Flow Diagram



OBJECTIVES



SCOPE & CONSTRAINTS



Constraints

- Data accuracy depends on bill image quality
- Initial insights limited by sample size
- Regulatory and privacy compliance requirements
- Dependence on partner adoption for monetization

CONCLUSION

This proposal presented a comprehensive study of a **Bill Scanning-Based Reward and Intelligence System**, addressing a real-world gap in how consumer purchase data is captured, utilized, and monetized. The report initially outlined the problem context, motivation, identified gaps, solution overview, system components, deployment model, objectives, scope, and constraints, clearly demonstrating the feasibility and relevance of the proposed system within a B.Sc. Computer Science study project framework.

Building upon this foundation, the extended technical design introduced a clear AI/ML-driven architecture, detailing the model types, data flow, and end-to-end machine learning pipeline from receipt capture and OCR, through data cleaning and categorization, to anonymized intelligence generation. The inclusion of dataset selection strategies, evaluation metrics, and a deeper literature review referencing both industry systems and academic benchmarks strengthens the project's academic rigor and aligns it with standard research and engineering practices.

Overall, the combined proposal establishes a scalable, privacy-aware, and technically grounded system that not only rewards users for data contribution but also transforms fragmented receipt data into meaningful, aggregated insights. The project is well-positioned for Phase-I implementation and validation, with clear pathways for future enhancement in predictive analytics, personalization, and real-time integrations, making it suitable for both academic evaluation and potential real-world extension.

Thanks..

Submission Date
02 | February | 2026

Team
ARAJ